Reference 1: JP 3-114100

S(N, m) indicates a size of spectrum of m frequency channels for n frames. The spectrum average value of the frames is derived from the following equation.

$$\bar{S} = 1/16 \sum_{l=0}^{15} S(n, l)$$

Here, $\bar{S}(n)$ indicates an average value of spectrum of the n-th frame. The spectral dispersion of this frame can be calculated by subtracting the spectrum average value from the spectrum of each frequency channel of the frame and squaring the subtraction result.

$$V(n) = \sum_{l=0}^{15} (S(n, l) - \bar{S}(n))^2$$

V(n) indicates the spectral dispersion of the n-th frame. FIG. 3 shows the spectrums of four inputted audio signals /d/, /s/, /a/, and /silence/. Out of the four audios, the silence spectrum is smoother than the other spectrums, so that the silence spectral dispersion is smaller than the other three audio spectral dispersions. Based on this characteristic, audio can be separated from background signals.

Reference numeral 71 is a spectral dispersion threshold calculator for calculating a spectral dispersion threshold by using spectral dispersions of some frames of an inputted audio, from the following equation.

$$VTH = (\sum_{n=1}^{15} V(n)) * 1.5$$

VTH here indicates a spectral dispersion threshold based on 10 frames, and V(n) indicates a spectral dispersion of the n-th frame within a silence period.

Reference numeral 81 is an audio clip detector for detecting an audio start point and an audio end point. By comparing the energy extracted by the energy extractor 20 and the spectral dispersion extracted by the spectral dispersion extractor 61 with the energy threshold calculated by the energy threshold calculator 30 and the spectral dispersion threshold calculated by the spectral dispersion threshold calculator 71, respectively, it is determined whether the frame is an audio start point.

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 03-114100
(43)Date of publication of application : 15.05.1991

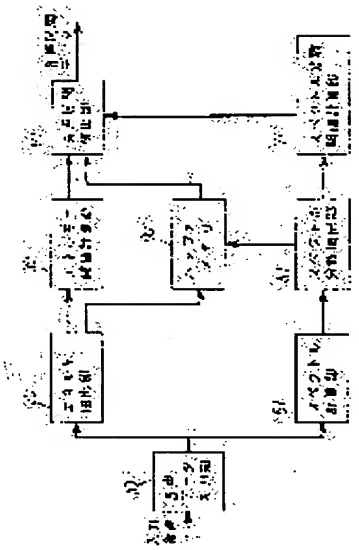
(51)Int.Cl. G10L 3/00
G10L 7/00
G10L 7/08

(21)Application number : 01-253313
(22)Date of filing : 28.09.1989

(71)Applicant : MATSUSHITA ELECTRIC IND CO LTD
(72)Inventor : TOU TOKUYOSHI
SO KINHIYOU
HAYASHI TOSHIHIRO

(54) VOICE SECTION DETECTING DEVICE

(57)Abstract:
PURPOSE: To detect even a voice of a frictional sound by detecting a start point and an end point of a voiced section of an input voice by a dynamic feature of a voice spectral dispersion and energy.
CONSTITUTION: The device is provided with an energy extracting part 20 for extracting the energy setting digital voice data in a prescribed section as one frame, an energy threshold calculating part 30 for adjusting an average value of background noise energy of a frame as an energy threshold, a spectral dispersion extracting part 61 for calculating a spectral dispersion by deriving an average value of a spectrum of the frame by a frequency of the frame, and a spectral dispersion threshold calculating part 71 for adjusting a spectral dispersion average value of a background noise as a spectral dispersion threshold. In this state, the energy and the spectral dispersion of each frame are compared with the energy threshold and the spectral dispersion threshold and whether it is a start point of a voice section or an end point is checked. In such a way, a voice start point of weak energy such as a frictional sound can be detected.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

⑲日本国特許庁(JP)　⑪特許出願公開

⑫公開特許公報(A)　平3−114100

⑤Int. Cl.⁵		識別記号	庁内整理番号
G 10 L	3/00	301 A	8842−5D
	7/00	A	8622−5D
	7/08	A	8842−5D

⑬公開　平成3年(1991)5月15日

審査請求　未請求　請求項の数 1　(全5頁)

⑭発明の名称　音声区間検出装置

⑳特　願　平1−253313
㉒出　願　平1(1989)9月28日

⑫発明者	鄧　徳淑	大阪府門真市大字門真1006番地	松下電器産業株式会社内
⑫発明者	蘇　錦標	大阪府門真市大字門真1006番地	松下電器産業株式会社内
⑫発明者	林　俊宏	大阪府門真市大字門真1006番地	松下電器産業株式会社内
⑪出願人	松下電器産業株式会社	大阪府門真市大字門真1006番地	
⑭代理人	弁理士 粟野重孝	外1名	

明　細　書

1、発明の名称

音声区間検出装置

2、特許請求の範囲

入力音声信号から一定区間内のディジタル音声データを一つのフレームとしてそのフレームでのエネルギーを抽出するエネルギー抽出部と、いくつかのフレームの背景雑音エネルギーを平均してその平均値をエネルギーしきい値とし、環境の変化に従って、エネルギーしきい値を調整するエネルギーしきい値計算部と、フレームの周波数によりまず該当フレームのスペクトルの平均値を求めてスペクトル分散をパラメータとして計算するスペクトル分散抽出部と、いくつかのフレームの背景雑音のスペクトル分散平均値をスペクトル分散しきい値として環境の変化に従って、適当にスペクトル分散しきい値を調整するスペクトル分散しきい値計算部と、入力音声の各フレームのエネルギー及びスペクトル分散をエネルギーしきい値及びスペクトル分散しきい値と比較することによりこのフレームは音声区間の始点であるか否かをチェックして、入力音声の各フレームのエネルギーをエネルギーしきい値と比較してこのフレームが音声区間の終点であるか否かをチェックする音声区間検出部とを備えたことを特徴とする音声区間検出装置。

3、発明の詳細な説明

産業上の利用分野

本発明は音声スペクトル分散(spectrum variance)とエネルギー(energy)との動態特徴(dynamic feature)によって入力音声の有声区間の始点と終点を検出し、音声認識システムに使用される音声区間検出装置である。

従来の技術

第2図は、従来の音声区間検出を示すブロック図である。同図において、10は、マイク、アナログ／ディジタル変換器(A/D converter) などの装置によって、アナログ音声信号を集めて、ディジタル音声データに変換して、バッファに記憶する音声データ入力手段である。20はフレームごと

にの各サンプルのパターンを２乗して、対数(log)を取り、このフレーム内のエネルギーとするエネルギー抽出手段である。30はいくつかのフレームの背景雑音(background noise)の雑音エネルギーを平均して、その平均値を音声検出用のエネルギーしきい値とするエネルギーしきい値獲得手段である。40はエネルギー抽出手段20のフレームでのエネルギーをエネルギーしきい値獲得手段30のエネルギーしきい値と比較して、フレームでのエネルギーはエネルギーしきい値より大きければ、このフレームは有声区間と見なされる。そして、もし有声区間の始点がまだ見付からない場合はこれは音声区間の始点と判定して、そうではない場合は音声区間内の継続音声と見なす音声始点検出手段である。50は音声始点を検出してから、エネルギー抽出手段20のフレームでのエネルギーをエネルギーしきい値獲得手段30のエネルギーしきい値と比較して、連続フレームのエネルギーはいくつかもエネルギーしきい値より小さい場合、このフレームは音声の終点と見なす音声終点検出手段である。

上述第２図に示す一般の音声区間検出は、例えば特公昭61-47437号公報（「音声区間の終端検出装置」）に上記のような類似方式を採用したものが示されている。また、特公昭61-3440号公報に示されている「音声区間信号検出回路」もゼロクロス率を基づいて上記のような類似方法で完成した装置である。

発明が解決しようとする課題

上記従来の音声区間検出は音声のエネルギー特徴だけにより音声区間の検出を行なう。しかしながら、中国語等の言語にはある音声の始点のエネルギーは非常に弱くて、雑音に含まれているエネルギーとよく似ている。例えば、ㄈ(f)、ㄙ(s)、ㄑ(q)、ㄘ(c)、などの摩擦音であれば従来の方法では雑音として誤判定して摩擦音の音声区間を正しく検出できない。それに従って、音声区間も乱れて、認識装置の認識率はもちろん大幅に低下する。

本発明はかかる点に鑑み、従来のエネルギー特徴からだけでは検出できない摩擦音等の音声も検出可能な音声区間検出装置を提供することを目的とする。

課題を解決するための手段

上記の問題点を解消するために、本発明は、入力音声信号から一定区間内のディジタル音声データを一つのフレームとしてそのフレームでのエネルギーを抽出するエネルギー抽出部と、いくつかのフレームの背景雑音エネルギーを平均してその平均値をエネルギーしきい値とし、環境の変化に従って、エネルギーしきい値を調整するエネルギーしきい値計算部と、フレームの周波数によりまず該当フレームのスペクトルの平均値を求めてスペクトル分散をパラメータとして計算するスペクトル分散抽出部と、いくつかのフレームの背景雑音のスペクトル分散平均値をスペクトル分散しきい値として環境の変化に従って、適当にスペクトル分散しきい値を調整するスペクトル分散しきい値計算部と、入力音声の各フレームのエネルギー及びスペクトル分散をエネルギーしきい値及びスペクトル分散しきい値と比較することによりこのフレームは音声区間の始点であるか否かをチェックして、入力音声の各フレームのエネルギーをエネルギーしきい値と比較してこのフレームが音声区間の終点であるか否かをチェックする音声区間検出部とを備えたことを特徴とする音声区間検出装置である。

作　用

本発明は上記した構成により、スペクトル分散手法で摩擦音のようなエネルギーの弱い音声始点をも検出できるので、確実に音声区間を検出することができる。

実施例

第１図は、本発明の一実施例を示すブロック図である。第１図において、第２図のものと同一動作を行なうものは、同一符号を付している。10は音声データ入力部、20はエネルギー抽出部、30はエネルギーしきい値計算部である。以上は第２図と同じ動作である。51は音声データ入力部10で入力された音声信号によって、各フレーム毎の線形

予測分析(ＬＰＣ、Linear Prediction Coding)関数でスペクトルを計算するスペクトル計算部、61はスペクトル平均値を計算してそしてスペクトル分散(Spectrum Variance)を計算するスペクトル分散抽出部である。90はエネルギー抽出部20で抽出した各フレームのエネルギーとスペクトル分散抽出部61で抽出した各フレームのスペクトル分散を記憶するバッファメモリである。

本実施例のサンプリング率(sampling rate)はたとえば10KHZで、各フレームは256ポイント、フレームとフレームとの間の重複部分は128ポイント、そして各フレーム毎の線形予測分析関数でスペクトルを計算する。スペクトルの計算はサンプリング・スペース(sampling space)の 0～5 KHzを16部分に分けて、各部分は一つの周波数チャンネル(channel)である。S(n,m)は第ｎ個フレームの第m個周波数チャンネルのスペクトルの大きさである。下式によりフレームのスペクトル平均値を獲得する。

$$\bar{S}(n)=1/16\sum_{i=0}^{16}S(n,\ i)$$

ここで、$\bar{S}(n)$は第ｎフレームのスペクトル平均値である。フレーム内の各周波数チャンネルのスペクトルからスペクトル平均値を引いてから２乗和してこのフレームのスペクトル分散を計算することができる。

$$V(n)=\sum_{i=0}^{16}(S(n,\ i)-\bar{S}(n))^2$$

Ｖ(ｎ)は第ｎ個フレームのスペクトル分散である。第３図は四つの入力音声信号/d/、/s/、/a/及び/silence/(沈黙)スペクトルである。この四つの音声に対して、沈黙のスペクトルは他のスペクトルよりスムーズなので、沈黙のスペクトル分散は他の三つの音声のスペクトル分散より小さい。この特性により音声を背景信号から分離することができる。

71は入力し始めた音声のいくつかのフレームのスペクトル分散により、下式のように計算してスペクトル分散しきい値となるスペクトル分散しきい値計算部である。

$$VTH=(\sum_{n=1}^{16}V(n))*1.5$$

ここでのＶＴＨは１０フレームに基づいてのスペクトル分散しきい値であり、Ｖ(ｎ)はSilence期間内の第ｎフレームのスペクトル分散である。

81は音声始点と音声終点を検出する音声区間検出部である。エネルギー抽出部20で抽出されたエネルギー及びスペクトル分散抽出部61で抽出されたスペクトル分散をそれぞれエネルギーしきい値計算部30でのエネルギーしきい値及びスペクトル分散しきい値計算部71でのスペクトル分散しきい値と比較して、そのフレームは音声始点であるかどうかをチェックする。すなわち、第４図の流れ図のように、まず一つのフレームのエネルギーを入力して、それのエネルギーがエネルギーしきい値より大きいかを判定して、もし、エネルギーしきい値より小さい場合は、このフレームを背景雑音と見なして、エネルギーしきい値を調整する。そうではない場合は、このフレームの後の５つのフレームがエネルギーしきい値より大きいかをチェックして、もし、大きければこのフレームは音声始点と一時的に暫定する。そして、このフレームの前の各フレームのスペクトル分散がスペクトル分散しきい値より大きいかどうかを、あるフレームのスペクトル分散がスペクトル分散しきい値より小さくなるまで検査する。そして、スペクトル分散がスペクトル分散しきい値より小さくなるなるフレームの直後のフレームを音声区間の始点とする。本当の音声始点はスペクトル分散がスペクトル分散しきい値より小さいフレームの後のこのフレームである。

第５図は、音声データの例を示す図で、この場合、まず連続する５フレームがエネルギーしきい値を越える第ｎフレームを音声始点と一時的に暫定する。次にこの第ｎフレームより前のフレームについて、スペクトル分散値を求め、スペクトル分散しきい値との比較を行なう。その結果、第ｎフレームから第ｎ－２フレームまではスペクトル

分散値がスペクトル分散しきい値より大きく、第ｎ－３フレームではスペクトル分散値がスペクトル分散しきい値より小さいので、第ｎ－２フレームを音声区間の始点とする。

音声終点の検出について、エネルギー抽出部20のフレームでのエネルギーをエネルギーしきい値と比較して、連続５フレームのエネルギーがいくつかエネルギーしきい値より小さい場合、このフレームは音声の終点と見なす。例えば中国語の場合、すべての音声の終点は母音なので、音声区間の終点の検出はエネルギーで判定することにより、正確に検出できる。

第５図においては、第ｍフレームから第ｍ＋５フレームまで連続する５つのフレームについて、そのエネルギーがエネルギーしきい値より小さいので、第ｍフレームを音声区間の終点とする。

上記本発明の実施例の各部で、音声のスペクトル分散の特性及びエネルギーにより、正確に音声の始点と終点を検出することができる。音声認識を行なう時に無駄な計算を減らすことができ、認識率も向上する。例えば、この音声区間の検出装置を特定話者に対する音声認識装置に適用し、中国語、都市名の１００単語の認識を行なった結果、認識率は９２％から９８％に向上し、認識時間の大幅な短縮も実現できた。

この発明は上記実施例に限定されることなく、その要旨を変更しない限り、適当に変化して実施することができる。例えば、周波数という特徴を獲得する方法は線形予測分析スペクトルに限らず、線形分析スペクトルや音声の周波数を表わす特徴でも適用できる。

また、データの入力方式はマイクに限らず、録音装置でも音声データをメモリに記憶させることができる。

また、この実施例では、エネルギーの比較を連続する５つのフレームについて行なっているが、このフレーム数については周辺の機器に応じて変更してやれば良い。さらに、この実施例で示したスペクトル分散しきい値ＶＴＨの式は実験値であり、機器の特性等に応じて各係数を調整すればより良い結果を得ることができる。

発明の効果

本発明は音声のエネルギー及びスペクトル分散の特性により有効に音声区間を検出することができる。そして、認識率を上げ、認識時間も大幅に短縮することができるのでその実用的効果は大きい。

4．図面の簡単な説明

第１図は本発明の一実施例における音声区間検出装置の構成を示すブロック図、第２図は従来例の音声区間検出装置の構成を示すブロック図、第３図は四つの入力音声信号のスペクトルを示す説明図、第４図は音声始点検出を説明する流れ図、第５図は音声区間の検出の例を示す説明図である。

10…音声データ入力部、20…エネルギー抽出部、

30…エネルギーしきい値計算部、

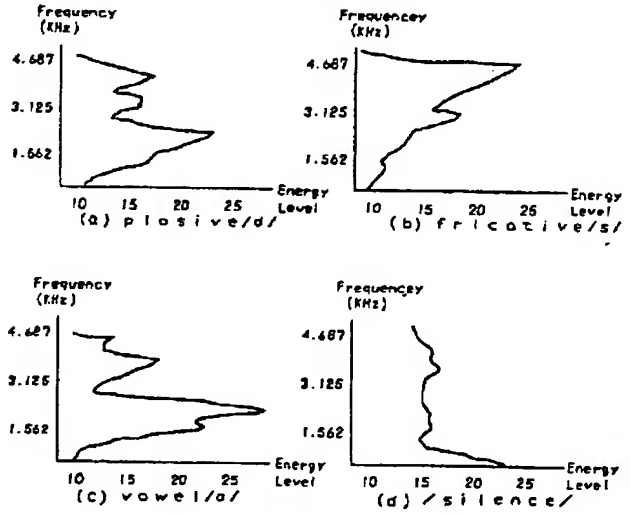
41…バッファメモリ、51…スペクトル計算部、

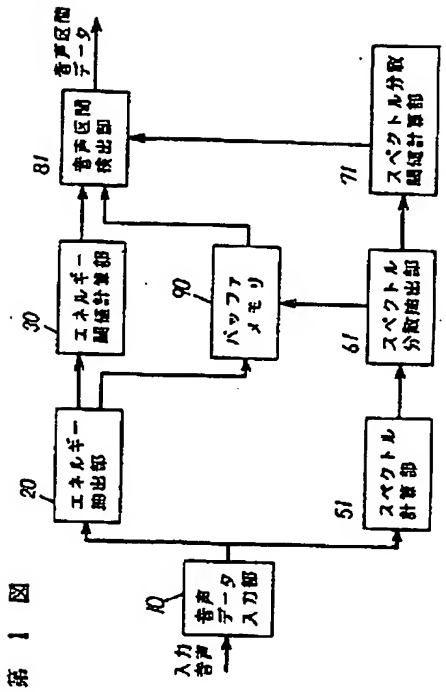
61…スペクトル分散計算部、

71…スペクトル分散しきい値計算部、

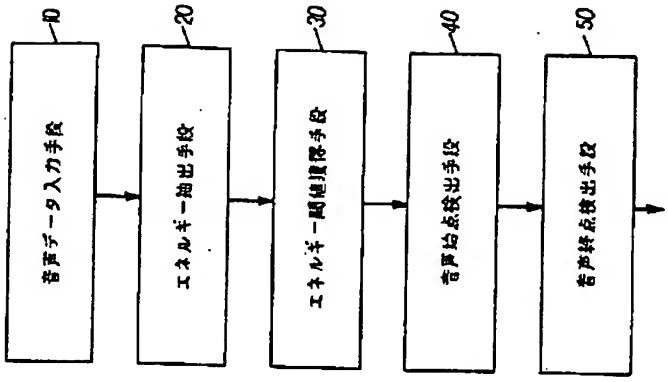
81…音声区間検出部。

第３図



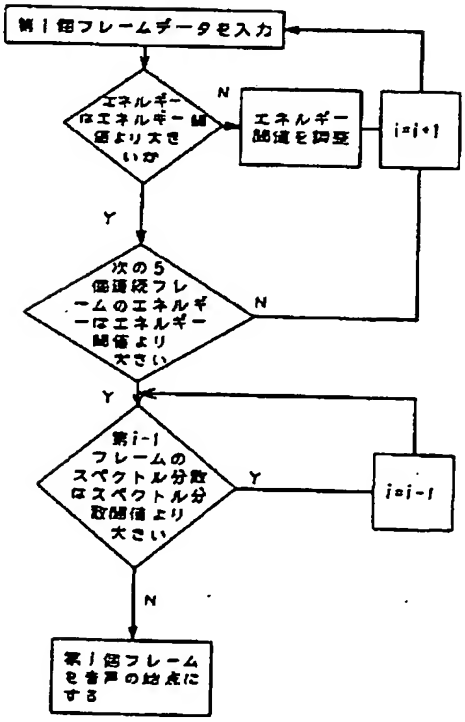


第 1 図



第 2 図

第 4 図



第 5 図